

Opportunistic Advertisement Scheduling in Live Social Media: A Multiple Stopping Time POMDP Approach

Vikram Krishnamurthy

Department of Electrical and Computer Engineering,

Cornell University, Ithaca, NY

vikramk@cornell.edu

Anup Aprem

Department of Electrical and Computer Engineering,

University of British Columbia, Vancouver, BC, Canada

aaprem@ece.ubc.ca

Sujay Bhatt

Department of Electrical and Computer Engineering,

Cornell University, Ithaca, NY

sh2376@cornell.edu

I. INTRODUCTION

Popularity of online video streaming has seen a sharp growth due to improved bandwidth for streaming and the ease of sharing User-Generated-Content (UGC) on the internet platforms. One of the primary motivations for users to generate content is that platforms like YouTube, Twitch etc., allow users to generate revenue through advertising and royalties. The revenue of Twitch which deals with live video gaming, play through of video games, and e-sport competitions, is around 3.8 billion for the year 2015, out of which 77% of the revenue was generated from advertisements¹.

Some of the common ways advertisements (ads) are scheduled on pre-recorded video contents on social media like YouTube are pre-roll, mid-roll and post-roll; where the names indicate the time at which the ads are displayed. In a recent research on viewer engagement, Adobe Research² concluded that mid-roll video ads constitute the most engaging ad type for pre-recorded video contents, outperforming pre-roll and post-rolls when it comes to completion rate (the probability that the ad will not be skipped). Viewers are more likely to engage with an ad if they are interested in the content of the video that the ad is been inserted into. Most content sharing platforms hosted *pre-recorded videos*, until recently, owing to higher

¹<http://www.tubefilter.com/2015/07/10/twitch-global-gaming-content-revenue-3-billion/>

²<https://gigaom.com/2012/04/16/adobe-ad-research/>

bandwidth requirements of real-time video streaming. However, this has changed lately with improved bandwidths (e.g. Google Fiber, Comcast Xfinity) and well known content sharing websites like YouTube and Facebook making provisions for *live streaming videos* to capture major events in real time³. Online live video streaming, popularly known as “Social TV”, now boasts a number of popular applications like Twitch, YouTube Live, Facebook Live, etc.

When a channel is streaming a *live video*, the mid-roll ads need to be scheduled manually⁶. Twitch allows only periodic ad scheduling [1] and YouTube and other live services currently offers no automated method of scheduling ads for live channels. The ad revenue in live channel depends on the click rate (the probability that the ad will be clicked), which in turn depend on the viewer engagement with the channel content. Hence, ads need to be scheduled when the viewer engagement of the channel is high. The problem of optimal scheduling of ads has been well studied in the context of advertising in television; see [2],[3], [4] and the references therein. However, scheduling ads on *live* online social media is different from scheduling ads on television in two significant ways [5]: (i) real-time measurement of viewer engagement (ii) revenue is based on ads rather than a negotiated contract. Prior literature on scheduling ads on social media is limited to ad scheduling in real-time for social network games, where the ads are served to either the video game consoles in real time over the Internet [6], or in digital games that are played via major social networks [7].

Problem Formulation: This paper deals with optimal scheduling of ads on live channels in social media, by considering viewer engagement, termed as *active scheduling*, to maximize the revenue generated from the advertisements. We model the viewer engagement of the channel using a Markov chain [8], [9]. The viewer engagement of the content is not observed directly, however, noisy observation of the viewer engagement is obtained by the current number of viewers of the channel. Hence, the problem of computing the optimal policy of scheduling ads on live channel can then be formulated as an instance of a stochastic control problem called the *partially observable Markov decision process* (POMDP). To the best of our knowledge, this is the first time in the literature that ad scheduling in live channels on social media is studied in a POMDP framework. The main contribution of this paper is two-fold:

- 1.) We provide a POMDP framework for the optimal ad-scheduling problem on live channels and show that it is an instance of the *optimal multiple stopping problem*.
- 2.) We provide structural results of the optimal policy of the multiple stopping problem and using stochastic approximation compute the best approximate policy.

Structural Results: The problem of optimal multiple stopping has been well studied in the literature see [10], [11], [12], [13] and the references therein. The optimal multiple stopping problem generalizes the classical (single) stopping problem, where the objective is to stop once to obtain maximum reward. Nakai [10] considers optimal L -stopping over a finite horizon of length N in a partially observed Markov chain. More recently, [13] considers L -stopping over a random horizon. The state of the finite horizon partially observed Markov chain in [10] above can be summarized by the “belief state”⁷. For a stopping time POMDP, the policy can be characterized by stopping region (set of belief state where we

³For example, as early as 2011, millions watched the Royal wedding live through YouTube⁴, and, more recently, the Caribbean Premier League is scheduled to be broadcast live on Facebook⁵

⁴<http://www.telegraph.co.uk/news/uknews/royal-wedding/8460801/Royal-wedding-Kate-and-William-marriage-live-on-YouTube.html>

⁵<http://www.si.com/tech-media/2016/07/06/caribbean-premier-league-matches-facebook-live>

⁶YouTube Live: Slate and Ad Insertion <https://is.gd/i6c7ku>

⁷The belief state is a sufficient statistic for all the past observations and actions [14].

stop) and continuance region (set of belief states where we do not stop). Nakai [10] shows that there are $N \times L$ stopping regions corresponding to each time index and stops and these regions form a nested structure. However, in live channels, the time horizon N is very large (in comparison to decision epochs) or initially unknown. Therefore, we extend the results in Nakai [10] to the infinite horizon case. The extension is both important and non-trivial. In the infinite horizon case, the policy is stationary (the stopping regions do not depend on the time index) and hence L stopping regions characterize the policy. We obtain similar structural results as [10] in the infinite horizon case.

Our main structural result (Theorem 1) is that the optimal policy is characterized by a switching curve on the unit simplex of Bayesian posteriors (belief states). To prove this result we use the monotone likelihood ratio stochastic order since it is preserved under conditional expectations. However, determining the optimal policy is non-trivial since the policy can only be characterized on a partially ordered set (more generally a lattice) within the unit simplex. We modify the MLR stochastic order to operate on line segments within the unit simplex of posterior distributions. Such line segments form chains (totally ordered subsets of a partially ordered set) and permit us to prove that the optimal decision policy has a threshold structure. Having established the existence of a threshold curve, Theorem 2 and Theorem 3 gives necessary and sufficient conditions for the best linear hyperplane approximation to this curve. Then a simulation-based stochastic approximation algorithm (Algorithm 1) is presented to compute this best linear hyperplane approximation.

Context: The optimal multiple stopping problem can be contrasted to the recent work on sampling with “causality constraints”. In sampling with causality constraints, not all the observations are observable. [15] considers the case where an agent is limited to a finite number of observations (sampling constraints) and must adaptively decide the observation strategy so as to perform quickest detection on a data stream. The extension to the case where the sampling constraints are replenished randomly is considered in [16]. In the multiple stopping problem, considered in this paper, there is no constraint on the observations and the objective is to stop L times at states that correspond to maximum reward.

The optimal multiple stopping problem, considered in this paper, is similar to sequential hypothesis testing [17], [18], sequential scheduling problem with uncertainty [19] and the optimal search problem considered in the literature. [20] and [21] consider the problem of finding the optimal launch times for a firm under strategic consumers and competition from other firms to maximize profit. However, in this paper we deal with sequential scheduling in a partially observed case. [22],[23] consider an optimal search problem where the searcher receives imperfect information on a (static) target location and decides optimally to search or interdict by solving a classical optimal stopping problem ($L = 1$). However, the multiple-stopping problem considered in this paper is equivalent to a search problem where the underlying process is evolving (Markovian) and the searcher needs to optimally stop $L > 1$ times to achieve a specific objective.

The paper is organized as follows: Section II provides a model of a live channel and introduces the notations, assumptions and key definitions. Section III provides the main results of the paper. First, similar to [10], we show that the stopping regions of the optimal ad scheduling policy form a nested structure. Second, we show the threshold structure of the optimal ad scheduling policy. In Section IV, we use the nested property of the stopping regions and the threshold property in Section III and stochastic approximation algorithm to compute the best approximate policies using linear thresholds. Such linear threshold policies are computationally inexpensive to implement. In Section V, we validate the model on three different datasets. First, we illustrate the analysis using a synthetic dataset and verify the performance of the optimal ad scheduling policy against conventional scheduling policies. Second, we show that the policy obtain by the multiple stopping problem can be used to detect changes in ground truth using data from online search. Finally, we use real datasets from

YouTube Live and Twitch to optimally schedule multiple ads ($L > 1$) in a sequential manner so as to maximize the revenue.

II. OPPORTUNISTIC SCHEDULING ON LIVE CHANNELS: MODEL AND PROBLEM FORMULATION

A. Live Channel Model

In this section, we develop a model of the live channel. The three main components of the live channel are i) Viewer engagement: How to model the viewer engagement of a live channel? ii) Dynamics of the channel viewers: How does the number of viewers vary with respect to the engagement? iii) Reward of the channel owner: What is the (monetary) reward obtained by the channel owner through advertising? Below, we develop models to address each of these questions.

- 1) **Dynamics of viewer engagement:** Similar to [24], viewer engagement, in the context of live channels, can be defined as the following process: (i) The viewer decides to watch the live channel. (ii) The viewer is “engaged” with the content of the live channel. (iii) The viewer will watch the live content without switching to other channels. (iv) The viewer is more attentive when watching the live content. The potential benefits of (iii) and (iv) above is that it increases the odds of the viewers being exposed and persuaded by advertisements.

Viewer engagement, as defined above, is an abstract concept which captures viewer attitude, behaviour and attentiveness. Archak et. al. [8], [9] developed a Markov chain model for online behaviour of users and the effects of advertising. The main finding is that the user behaviour in online social media can be approximated by a first-order Markov chain. Following Archak et. al. [8], [9], we model the viewer engagement at time t , denoted by X_t , as an S -state Markov chain with state-space $\mathcal{S} \equiv \{1, 2, \dots, S\}$. The dynamics of the viewer engagement of the channel, modelled as a Markov chain, can be characterized by the transition matrix P and initial probability vector π_0 as follows:

$$P(i, j) = \mathbb{P}(X_{t+1} = j | X_t = i), \quad \pi_0(i) = \mathbb{P}(X_0 = i) \quad (1)$$

The Markov chain model for the viewer engagement of the channel is validated using simulations in Sec. V.

- 2) **Dynamics of channel viewers:** The number of viewers at time t depend on the viewer engagement of the live channel. As viewers are more engaged with the content, they are less likely to switch the channel. Hence, a higher viewer engagement state has higher number of viewers compared to a lower engagement state. Therefore, we model the dynamics of channel viewers as follows: The number of viewers at time t , denoted by Y_t , belongs to the countably infinite set \mathcal{Y} of non-negative integers. Denote, the conditional probability of j viewers ($Y_t = j$) in viewer engagement state i ($X_t = i$) by $B(i, j)$. Note that the conditional probability $B(i, j)$ is assumed to be time homogeneous⁸. The number of viewers Y_t is modeled as a Poisson random variable with state dependent mean $g_i, i \in \mathcal{S}$, based on evidence in [25] and [11], as follows:

$$B(i, j) = \mathbb{P}(Y_t = j | X_t = i) = \frac{g_i^j \exp(-g_i)}{j!}, \quad \forall i \in \mathcal{S}, j \in \mathcal{Y}. \quad (2)$$

The states with higher state dependent mean correspond to states with higher viewer engagement.

The channel owner does not observe the true viewer engagement of the channel, X_t . However, at each time instant t , the channel owner receives a noisy observation of the viewer engagement of the channel by the number of viewers, Y_t . Hence, the channel owner needs to estimate the viewer engagement using the history of noisy observations and schedule ads.

⁸The conditional probability $B(i, j)$ does not depend on the time index, t .

3) **Reward of channel owner:** The channel owner agrees to show L ads during the live session, which are decided prior to the beginning of the session. For example, the ads during the Super Bowl 50 in YouTube Live had to be pre-booked in advance.⁹ Hence, at each time instant t , the channel owner chooses an action u_t as follows: The channel owner can continue with the live session (denoted by *Continue*) or can pause the live session to insert an ad (denoted by *Stop*). Hence, $u_t \in \mathcal{A} = \{\text{Stop}(1), \text{Continue}(2)\}$ ¹⁰ denote the actions available to the channel owner at time t . This problem of scheduling the L ads opportunistically, so as to obtain maximum revenue, corresponds to a multiple stopping problem with L stops.

Choosing to stop at time t (and schedule an ad), with l stops remaining (l ads remaining), the channel owner will accrue a reward $r_l(X_t, a = 1)$, where X_t is the state of the Markov chain at time t . The reward obtained by the channel owner depends on two factors: (i) the number of viewers (ii) the completion rate and click rate of the ad. To capture these, we model the reward as follows:

$$r_l(X_t = i, a = 1) = \alpha_i g_i. \quad (3)$$

In (3), g_i captures the average number of viewers in any viewer engagement state. The term $\alpha_i \in [0, 1]$ captures the completion and click rate of the ads at any viewer engagement state. Similarly, if the channel owner chooses to continue, he will accrue $r_l(X_t, a = 2)$. When an ad is not shown, the reward obtained by the channel owner is usually zero, hence, $r_l(X_t, a = 2) = 0$.

Hence, to maximize revenue, the channel owner needs to opportunistically schedule ads at time slots when the viewer engagement is high, corresponding to a higher number of viewers and higher click rate.

B. Ad Scheduling : Problem formulation & Stochastic Dynamic Programming

1) *Problem Formulation:* The ad scheduling policy (or the control policy) prescribes a decision rule that determines the action taken by the channel owner. Let the initial probability vector, π_0 and the history of past observations at time t for the channel owner be denoted as $Z_t = \{\pi_0, Y_1, \dots, Y_t\}$. The control policy, at time t , maps the history Z_t to action. Hence, the policy of the channel owner μ belongs to the set of admissible policies $\mathcal{U} = \{\mu : \mu \text{ maps } Z_t \rightarrow \mathcal{A}\}$. Below, we reformulate the sequential multiple stopping problem of scheduling ads in terms of belief state.

Let $\Pi = \{\pi \in \mathbb{R}^S : \mathbf{1}'_S \pi = 1, \pi(i) \geq 0\}$ denote the belief space of all S -dimensional probability vectors. The belief-state π_t is a sufficient statistic of Z_t [14], and evolves as $\pi_{t+1} = T(\pi_t, Y_{t+1})$, where

$$T(\pi_t, Y_{t+1}) = \frac{B_{Y_{t+1}} P' \pi_t}{\sigma(\pi_t, Y_{t+1})}, \quad \sigma(\pi_t, Y_{t+1}) = \mathbf{1}'_S B_{Y_{t+1}} P' \pi_t, \quad (4)$$

$$B_{Y_{t+1}} = \text{diag}(B(1, Y_{t+1}), \dots, B(S, Y_{t+1})).$$

Here $\mathbf{1}_S$ represents the S -dimensional vectors of ones.

The aim is to compute the optimal stationary ad scheduling policy $u_t = \mu(\pi_t, l)$, as a function of the belief, π_t , and the number of stops (or the number of ads) remaining, l , to maximize the infinite horizon criterion defined in (6). Let τ_l denote the stopping time when there are l stops remaining:

$$\tau_l = \inf \{t : t > \tau_{l+1}, u_t = 1\}, \text{ with } \tau_{L+1} = 0. \quad (5)$$

⁹<http://www.campaignlive.co.uk/article/youtube-launches-real-time-ads-major-live-events-starting-super-bowl-50/1380260>

¹⁰The Stop and Continue actions will be denoted by 1 and 2, respectively in the remainder of the paper.

The infinite horizon discounted criterion with stationary policy μ , and initial belief π_0 is as below:

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau_L-1} \rho^t r_L(X_t, 2) + \rho^{\tau_L} r_L(X_{\tau_L}, 1) + \sum_{t=\tau_L+1}^{\tau_{L-1}-1} \rho^t r_{L-1}(X_t, 2) + \cdots + \rho^{\tau_1} r_1(X_{\tau_1}, 1) \mid \pi_0 \right\}, \quad (6)$$

$$= \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau_L-1} \rho^t r'_{2,L} \pi + \rho^{\tau_L} r'_{1,L} \pi + \sum_{t=\tau_L+1}^{\tau_{L-1}-1} \rho^t r'_{2,L-1} \pi + \cdots + \rho^{\tau_1} r'_{1,1} \pi \mid \pi_0 \right\}, \quad (7)$$

where $r_{u,l} = [r_l(1, u), \dots, r_l(S, u)]'$. In (6) and (7), $\rho \in (0, 1]$ denotes the discount factor. Below, we study the special case, where $r_{1,1} = r_{1,2} = \cdots = r_{1,L} = r$ and $r_{2,1} = r_{2,2} = \cdots = r_{2,L} = 0$, i.e. the reward for stopping and scheduling an ad is independent of the number of stops remaining and the channel owner accrues no reward for continuing. The extension to the general case is straightforward.

2) *Stochastic Dynamic Programming*: The computation of the optimal ad scheduling policy μ^* , to maximize the infinite horizon discounted criterion in (6) and (7), is equivalent to solving Bellman's dynamic programming equation [14]:

$$\mu^*(\pi, l) = \operatorname{argmax}_{u \in \mathcal{A}} Q(\pi, l, u), \quad V(\pi, l) = \max_{u \in \mathcal{A}} Q(\pi, l, u), \quad (8)$$

where,

$$Q(\pi, l, 1) = r' \pi + \rho \sum_{Y \in \mathcal{Y}} V(T(\pi, Y), l-1) \sigma(\pi, Y), \quad Q(\pi, l, 2) = \rho \sum_{Y \in \mathcal{Y}} V(T(\pi, Y), l) \sigma(\pi, Y) \quad (9)$$

The above dynamic programming formulation is a POMDP. Since the state-space Π , is a continuum, Bellman's equation (8) does not translate into practical solution methodologies as $V(\pi, l)$ needs to be evaluated at each $\pi \in \Pi$. This in turn renders the calculation of the optimal policy $\mu^*(\pi, l)$ computationally intractable. In Sec. III we derive structural results of the optimal ad scheduling policy. The advantage of deriving structural results is that the optimal policy can be computed efficiently. Sec. IV provides stochastic approximation algorithms to compute approximations of the optimal policy using the structural results derived in Sec. III.

III. OPTIMAL AD SCHEDULING: STRUCTURAL RESULTS

In this section, we derive structural results for the optimal ad scheduling policy. We first introduce the value iteration algorithm in Sec. III-A, a successive approximation method to solve the dynamic programming recursion in (8). The value iteration algorithm is a valuable tool for deriving the structural results. In Section III-B, we use the value iteration algorithm in Sec. III-A to prove structural results of the optimal ad scheduling policy. Using the structural results in Sec. III-B, we provide a simulation based algorithm to compute the policy in Sec. IV.

A. Value Iteration Algorithm

The value iteration algorithm is a successive approximation approach for solving Bellman's equation (8).

The procedure is as follows: For iterations $k = 0, 1, \dots$, the value function $V_k(\pi, l)$ and the policy $\mu_k(\pi, l)$ is obtained as follows

$$V_{k+1}(\pi, l) = \max_{u \in \{1,2\}} Q_{k+1}(\pi, l, u), \quad (10)$$

$$\mu_{k+1}(\pi, l) = \operatorname{argmax}_{u \in \{1,2\}} Q_{k+1}(\pi, l, u), \quad (11)$$

where

$$Q_{k+1}(\pi, l, 1) = r' \pi + \rho \sum_y V_k(T(\pi, y), l-1) \sigma(\pi, y), \quad (12)$$

and

$$Q_{k+1}(\pi, l, 2) = \rho \sum_y V_k(T(\pi, y), l) \sigma(\pi, y), \quad (13)$$

with $V_0(\pi, l)$ initialized arbitrarily. It can be easily shown that the above procedure converges [14].

In order to prove the structural results of the stationary ad scheduling policy, defined in Sec. II, we define the stopping and the continuance regions of the policy as below. Let $W_k(\pi, l)$ be defined as

$$W_k(\pi, l) \triangleq V_k(\pi, l) - V_k(\pi, l-1). \quad (14)$$

The stopping and continuance region (at each iteration k) is defined as follows:

$$\begin{aligned} S_{k+1}^l &= \{\pi | r'\pi \geq \rho \sum_y W_k(T(\pi, y), l) \sigma(\pi, y)\} \\ C_{k+1}^l &= \{\pi | r'\pi < \rho \sum_y W_k(T(\pi, y), l) \sigma(\pi, y)\} \end{aligned} \quad (15)$$

Since the value iteration converges, the optimal stationary policy $\mu^*(\pi, l)$ is defined as

$$\mu^*(\pi, l) = \lim_{k \rightarrow \infty} \mu_k(\pi, l). \quad (16)$$

Correspondingly, the stationary stopping and continuance sets are defined by

$$S^l = \lim_{k \rightarrow \infty} S_k^l, \quad C^l = \lim_{k \rightarrow \infty} C_k^l. \quad (17)$$

The value function, $V_k(\pi, l)$ in (10), can be rewritten, using (15), as follows:

$$V_k(\pi, l) = \left(r'\pi + \rho \sum_y V_{k-1}(T(\pi, y), l-1) \sigma(\pi, y) \right) \mathcal{I}_{S_k^l} + \left(\rho \sum_y V_{k-1}(T(\pi, y), l) \sigma(\pi, y) \right) \mathcal{I}_{C_k^l} \quad (18)$$

where $\mathcal{I}_{C_k^l}$ and $\mathcal{I}_{S_k^l}$ are indicator functions on the continuance and stopping regions respectively, for each iteration k .

Assume $S_k^{l-1} \subset S_k^l$ (see Proposition 3) and substituting (18) in the definition of $W_k(\pi, l)$ in (14),

$$\begin{aligned} W_k(\pi, l) &= \left(\rho \sum_y W_{k-1}(T(\pi, y), l) \sigma(\pi, y) \right) \mathcal{I}_{C_k^l}(\pi) \\ &\quad + r'\pi \mathcal{I}_{C_k^{l-1} \cap S_k^l}(\pi) \\ &\quad + \left(\rho \sum_y W_{k-1}(T(\pi, y), l-1) \sigma(\pi, y) \right) \mathcal{I}_{S_k^{l-1}}(\pi) \end{aligned} \quad (19)$$

The value iteration algorithm (10) to (13) does not translate into a practical solution methodology since the belief state belongs to an uncountable set. Hence, there is a strong motivation to characterize the structure of the optimal policy. In the following section, we use the definition of $W_k(\pi, l)$ in (19) to prove structural results on the stopping region defined in (15).

B. Structural results for the optimal ad scheduling policy

In this section, we provide structural results on the optimal policy of the multiple stopping problem corresponding to maximizing the ad revenue.

1) *Assumptions:* The main result below, namely, Theorem 1, requires the following assumptions on the reward vector, r , and the transition matrix, P and the observation distribution, B . The first assumption on the reward says that e_1 is the state with the highest reward and the reward monotonically decreases. This captures the channel owners' preference of scheduling ads at the highest viewer engagement state. A sufficient condition for the reward to monotonically decrease is that both the mean number of viewers g_i , and the completion and click rate α_i be non-increasing. The second and the third assumptions (A2) and (A3) relate to the underlying stochastic model and is related to MLR ordering of the updated belief vector in (4) (see Theorem 4 and Theorem 5 in Appendix A). The assumption (A2) models the following facts: i) The user behaviour in online social media can be approximated by a first order Markov chain. ii) The viewer engagement changes at a smaller time scale compared to sampling or decision epochs. The assumption (A3) is due to the fact that the viewer engagement states can be ordered corresponding to decreasing mean number of viewers, g_i . The last assumption (A4) is a technical condition required for our proof.

- (A1) The vector, r , has decreasing elements. Hence, $r'\pi$ is increasing in π .
- (A2) B is TP2¹¹. A necessary and sufficient condition for B to be TP2 is that the state dependent mean g_i in (2) is monotonically decreasing.
- (A3) P is TP2¹¹.
- (A4) The vector, $(I - \rho P')r$, has decreasing elements.

2) *Main Result:* The following (Theorem 1) is the main result of the paper and the proof is provided in Appendix D. Theorem 1.A states that the optimal policy is a monotone policy. The optimal policy $\mu^*(\pi, l)$ decreases monotonically with the belief state π . However, for a monotone policy to be well defined, we need to first define the ordering between two belief states. In this paper, we use the Monotone Likelihood Ratio (MLR) ordering¹², and the less restrictive MLR ordering on lines $\mathcal{L}(e_1, \bar{\pi})$ and $\mathcal{L}(e_S, \bar{\pi})$ ¹³ over the belief states [26], [27]. Fig. 1 illustrates the definition of $\mathcal{L}(e_1, \bar{\pi})$.

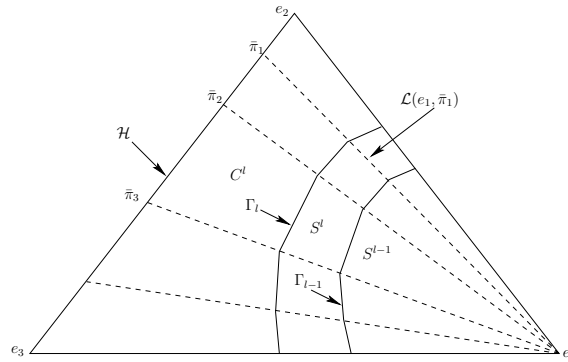


Fig. 1. $\mathcal{L}(e_1, \bar{\pi})$ corresponds to a line joining e_1 and any $\bar{\pi} \in \mathcal{H}$ ¹⁵ on the simplex Π . The advantage of MLR ordering on lines is that the belief states on the line $\mathcal{L}(e_1, \bar{\pi})$ and $\mathcal{L}(e_S, \bar{\pi})$ can be fully ordered. Hence, we can define monotonicity of the policy over the lines. This is not possible for MLR ordering, since it is a partial order.

¹¹Refer to Appendix A for the definition of TP2 ordering.

¹²MLR ordering is defined in Def. 2 in Appendix B

¹³MLR ordering over lines is defined in Appendix C

¹⁵ \mathcal{H} is defined in Appendix C

Theorem 1: Assume (A1), (A2), (A3) and (A4), then,

- A There exists an optimal policy $\mu^*(\pi, l)$ that is monotonically decreasing on lines $\mathcal{L}(e_1, \bar{\pi})$, and $\mathcal{L}(e_X, \bar{\pi})$ for each l .
- B There exists an optimal switching curve Γ_l , for each l , that partitions the belief space $\Pi(X)$ into two individually connected regions S^l and C^l , such that the optimal policy is

$$\mu^*(\pi, l) = \begin{cases} 1 & \text{if } \pi \in S^l \\ 2 & \text{if } \pi \in C^l \end{cases} \quad (20)$$

- C $S^{l-1} \subset S^l$.

Theorem 1A implies that the optimal action is monotonically decreasing on the line $\mathcal{L}(e_1, \bar{\pi})$, as shown in Fig. 1. Hence, on each line $\mathcal{L}(e_1, \bar{\pi})$ there exists a threshold above (in MLR sense) which it is optimal to *Stop* and below which it is optimal to *Continue*. Theorem 1B says, for each l , the stopping and continuance sets are connected. Hence, there exists a threshold curve, Γ_l , as shown in Fig. 1, obtained by joining the thresholds, from Theorem 1A, on each of the line $\mathcal{L}(e_1, \bar{\pi})$. Theorem 1C proves the sub-setting structure of the stopping and continuance sets, as shown in Fig. 1.

In order to prove the theorem, we introduce the following propositions, proofs of which are provided in the Appendix D. The value function of the classical (single) stopping POMDP increases with π (in MLR sense) [27], [28], [29]. Proposition 1 states that the above result holds even in the multiple stopping problem. In the classical stopping POMDPs in [27], the stopping and continuance sets are characterized using the value function. However, in the multiple stopping problem, considered in this paper, W plays the role of value function in characterizing the stopping and continuance region. Proposition 2 proves the corresponding result in the multiple stopping problem. Proposition 3 proves the nested stopping regions at each iteration k of the value iteration. Since the value iteration converges, Proposition 3 implies Theorem 1C.

Proposition 1: $V_k(\pi, l)$ is increasing in π .

Proposition 2: $W_k(\pi, l)$ is decreasing in l .

Proposition 3: $S_{k+1}^l \supset S_{k+1}^{l-1}$

To summarize, we showed the following properties of the optimal policy:

- (i) the optimal policy is monotone on the lines $\mathcal{L}(e_1, \pi)$.
- (ii) existence of a unique threshold curve for the stopping region, Γ^l .
- (iii) the stopping regions have a sub-setting property, i.e. $S^{l-1} \subset S^l$.

IV. STOCHASTIC APPROXIMATION ALGORITHM FOR COMPUTING BEST APPROXIMATE AD SCHEDULING POLICY

In this section, we synthesize policies satisfying the properties of the optimal ad scheduling policy derived in Sec. III. The policies can be characterized by L threshold curve, corresponding to each of the stopping regions. In this section, we parametrize the threshold curve, Γ^l , as $\hat{\Gamma}_\theta^l$. Here, θ denotes the parameter of the threshold curve. To capture the essence of Theorem 1, we require that the parametrized optimal policy, $\mu_\theta(l, \pi)$, be decreasing on lines $\mathcal{L}(e_1, \bar{\pi})$ and $\mathcal{L}(e_S, \bar{\pi})$ and the stopping sets are connected and satisfy the sub-setting property, i.e. $S^{l-1} \subset S^l$. Below, we will restrict our attention to obtain the best *linear* threshold policy, i.e. policy of the form given in (21). We characterize the parameters of the threshold policy in Sec. IV-A and provide an algorithm to compute the parameters using stochastic approximation in Sec. IV-B.

A. Structure of best linear MLR threshold policy for ad scheduling

Consider a threshold hyperplane, on the simplex Π , of the form (21) where $\theta_l \in \mathbb{R}^{S-1}$ denotes the coefficient vector. The linear threshold scheduling policy, denoted by $\mu_\theta(l, \pi)$ is defined as

$$\mu_\theta(l, \pi) = \begin{cases} 1 & \text{if } \begin{bmatrix} 0 & 1 & \theta_l \end{bmatrix} \begin{bmatrix} \pi \\ -1 \end{bmatrix} \leq 0 \\ 2 & \text{else,} \end{cases} \quad (21)$$

Here, $\theta = (\theta_1, \theta_2, \dots, \theta_L) \in \mathbb{R}^{L \times (S-1)}$ is the concatenation of the θ_l vectors.

To capture the essence of Theorem 1A, we require that the policy be decreasing on lines, i.e. for $\pi_1 \geq_{\mathcal{L}_1} \pi_2$, $\mu_\theta(\pi_1, l) \leq \mu_\theta(\pi_2, l)$. Theorem 2 gives necessary and sufficient conditions on the coefficient vector θ_l such that the above condition holds.

Theorem 2: A necessary and sufficient condition for the linear threshold policy $\mu_\theta(l, \pi)$ to be

- 1) MLR decreasing on line $\mathcal{L}(e_1, \bar{\pi})$, iff $\theta_l(S-1) \geq 0$ and $\theta_l(i) \geq 0$, $i \leq S-2$.
- 2) MLR decreasing on line $\mathcal{L}(e_S, \bar{\pi})$, iff $\theta_l(S-1) \geq 0$, $\theta_l(S-2) \geq 1$ and $\theta_l(i) \leq \theta_l(S-2)$, $i < S-2$.

The proof of Theorem 2 is similar to the proof of Theorem 12.4.1 in [27] and hence is omitted in this paper.

The stopping sets are connected since we parametrize the threshold curve using a linear hyperplane. Finally, the linear threshold approximation curve needs to satisfy the sub-setting property in Theorem 1C. Theorem 3 provides sufficient conditions such that the parametrized linear threshold curve satisfy the sub-setting property and the proof is provided in the Appendix I.

Theorem 3: To satisfy the sub-setting structure in Theorem 1C, the parameters of the linear threshold curve have to satisfy the following condition

$$\begin{aligned} \theta_{l-1}(S-1) &= \theta_l(S-1) \\ \theta_{l-1}(i) &\geq \theta_l(i) \quad i < S-1 \end{aligned} \quad (22)$$

Therefore, under the conditions of Theorem 2 and Theorem 3 the linear threshold policy in (21) satisfy all the conditions in Theorem 1 and hence qualify as the *best* linear threshold policy.

The parameter θ can be re-parametrized as follows:

$$\theta_l^\phi(i) = \begin{cases} \phi_1^2(S-1) & i = S-1 \\ 1 + \phi_1^2(S-2) \prod_{\ell=2}^l \sin^2(\phi_\ell(S-2)) & i = S-2 \\ \left(1 + \phi_1^2(S-2) \prod_{\ell=2}^l \sin^2(\phi_\ell(S-2))\right) \sin^2(\phi_l(i)) & i < S-2 \end{cases} \quad (23)$$

It can be easily checked the parametrization in (23) satisfies the conditions in Theorem 2 and Theorem 3.

Theorem 2 and Theorem 3 characterize the parameters of the linear threshold policy. In Sec. IV-B we provide an algorithm to compute the best linear threshold policy satisfying Theorem 2 and Theorem 3.

B. Simulation-based stochastic approximation algorithm for estimating the best linear MLR threshold policy for ad scheduling

In this section, we provide an algorithm to compute the best linear thresholds satisfying the conditions in Theorem 2 and Theorem 3. Recall, the optimal policy minimizes the average discounted reward in (6) and (7). The optimal linear

thresholds can be obtained by a policy gradient algorithm [14]. Algorithm 1 is a policy gradient algorithm to compute the best linear threshold policy. In this algorithm, we approximate J_μ in (6) and (7) by the finite time approximation

$$J_N(\theta) = \mathbb{E} \left\{ \sum_{l=1}^L \rho^{\tau_l} r(X_{\tau_l}, 1) \mid \tau_l \leq N; \forall l \right\}. \quad (24)$$

Here, we have made explicit the dependence of the parameter vector, θ , on the discounted reward and with an abuse of notation, have suppressed the dependence on the policy μ . It can be shown that $J_N(\theta)$ is an asymptotically biased estimate of J_μ and can be obtained by simulation for large N .

Algorithm 1 Threshold-Based Policy Gradient Algorithm for Optimal Multiple Stopping

Require: Assume the parameters of the optimal multiple stopping problem satisfy the assumptions in Theorem 1.

- 1: Initialize: Choose initial parameters $\hat{\phi}_0$ and initial linear threshold policy $\mu_{\hat{\theta}_0}$ using (21).
 - 2: **for** each iterations $n = 0, 1, 2, \dots$: **do**
 - 3: Evaluate cost $J_N(\theta^{\hat{\phi}_n})$ using (24) and gradient estimate $\nabla_\phi J_N(\theta^{\hat{\phi}_n})$ with policy $\mu_{\theta^{\hat{\phi}_n}}$ using (25).
 - 4: Update the parameter vector $\hat{\phi}_n$ to $\hat{\phi}_{n+1}$ using (27).
 - 5: **end for**
-

The policy gradient algorithm in Algorithm 1 requires the gradient $\nabla_\phi J_N(\cdot)$ at each iteration. Computation of the gradient is quite difficult due to the non-linear dependence of the parameter ϕ on the cost function. Hence, in this paper, we estimate the gradient through a stochastic approximation algorithm.

There are several stochastic approximation algorithms available in the literature: infinitesimal perturbation analysis[30], weak derivatives [31] and the SPSA algorithm [32]. In this paper, we use the SPSA algorithm and because of the constraints in Theorem 3 we use a variant of SPSA that can handle linear inequality constraints [33].

Following [32] and [33], the gradient estimate is obtained by picking a random direction ω_n , at each iteration n . The estimate of the gradient is then given by

$$\hat{\nabla}_\phi J_N(\theta^{\hat{\phi}_n}) = \frac{J_N(\theta^{\hat{\phi}_n + c_n \omega_n}) - J_N(\theta^{\hat{\phi}_n - c_n \omega_n})}{2c_n} \omega_n, \quad (25)$$

where,

$$\omega_n(i) = \begin{cases} -1 & \text{with probability 0.5} \\ +1 & \text{with probability 0.5.} \end{cases} \quad (26)$$

The equations for the parameter update are as follows [33]:

$$\phi_{n+1} = \phi_n - a_n \hat{\nabla}_\phi J_N(\theta^{\hat{\phi}_n}). \quad (27)$$

The parameters a_n and c_n and r_n are chosen as in [33] as follows:

$$\begin{aligned} a_n &= \varepsilon(n+1+\varsigma)^{-\kappa} & 0.5 < \kappa \leq 1, \quad \text{and } \varepsilon, \varsigma > 0 \\ c_n &= \mu(n+1)^{-\Upsilon} & \mu > 0 \end{aligned} \quad (28)$$

The stochastic approximation in Algorithm 1 converges to a local minimum. Hence, it is necessary to try several initial conditions and pick the best threshold.

V. NUMERICAL RESULTS: SYNTHETIC AND REAL DATASET

In this section, we present numerical results on synthetic and real datasets. First, we illustrate our analysis of Theorem 1, using synthetic data. Second, we demonstrate how the optimal multiple stopping framework, used for ad scheduling in live media, can be used to detect changes in ground truth using real data from online search. Online search is linked to advertising in television and online social media [34]. Third, we show, through simulations, that the scheduling policy obtained from Algorithm 1 outperforms conventional technique of scheduling ads in live social media. In this paper, we compare the scheduling policy obtained from Algorithm 1 with two schemes: “Periodic” and “Random”. The periodic scheme models the most common method of advertisement scheduling in pre-recorded videos in platforms like YouTube. In the context of live channels, Twitch uses a periodic scheduling where an ad is inserted periodically into the live channel [1]. In contrast, in the random scheduling scheme, the ad is inserted randomly into the live channel. The random scheduling scheme is used as a benchmark to compare the revenue obtained through the periodic scheme and the policy obtained through Algorithm 1.

A. Synthetic Data

In this section, we do a simulation study based on synthetic data for the optimal multiple stopping problem. The objective is to illustrate the analysis in Theorem 1. In addition, we obtain the policy by solving the dynamic programming equations (8) and show through simulations that it outperforms conventional method of scheduling ads periodically.

In this “toy” example, the viewer engagement of the live channel can be categorized into 3 states: “Popular”, “Interesting” and “Boring”, denoted by 1, 2 and 3, respectively. The transition between the various states follow a Markov chain with transition matrix given in (29). The viewer engagement state is observed through a state dependent Poisson process with mean given in (30). The reward structure for scheduling the ads is as in (31).

$$P = \begin{bmatrix} 0.2 & 0.1 & 0.7 \\ 0.1 & 0.1 & 0.8 \\ 0 & 0.1 & 0.9 \end{bmatrix} \quad (29)$$

$$g = \begin{bmatrix} 12 & 7 & 2 \end{bmatrix} \quad (30)$$

$$r = \begin{bmatrix} 9 & 3 & 1 \end{bmatrix} \quad (31)$$

Fig. 2 shows a sample trajectory of the observation and the state sequence. For $L = 5$, i.e. for a total of 5 ads to be scheduled, we obtained the policy by solving the dynamic programming equations in (8). The resulting policy, in terms of stopping and continuance set as defined in (14), is shown in Fig. 3. We also show the corresponding points where the policy was chosen in Fig. 2. Only the stopping set for $l = 5$ and $l = 1$, i.e. S^5 and S^1 respectively are shown in Figure 3. As can be seen from Figure 3, $S^1 \subset S^5$, verifying Theorem 1.C. The stopping regions S^1 and S^5 are connected and the optimal policy is threshold on any line $\mathcal{L}(e_1, \pi)$, verifying Theorem 1.B and Theorem 1.A respectively.

We now compare the policy obtained by solving the dynamic programming equation in (8) with the conventional technique of scheduling ads periodically within the live session. We also include a random scheduling policy, where the ads are scheduled randomly within the live session, for benchmarking purposes. Fig. 4 shows the comparison between the various schemes. The results in Fig. 4 was obtained by 10^4 independent Monte Carlo (M.C.) simulations. It can be seen from Fig. 4 that the policy obtained from (8) significantly outperforms (close to 4 times) periodic scheduling. This is not surprising since the policy obtained by solving the dynamic programming in (8) opportunistically schedules ads when the viewer engagement of the channel is high.

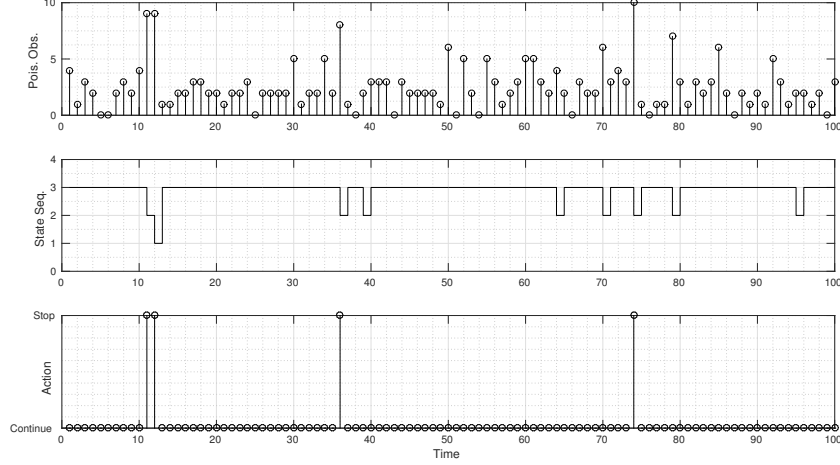


Fig. 2. Sample trajectory of the state, observation sequence and the policy.

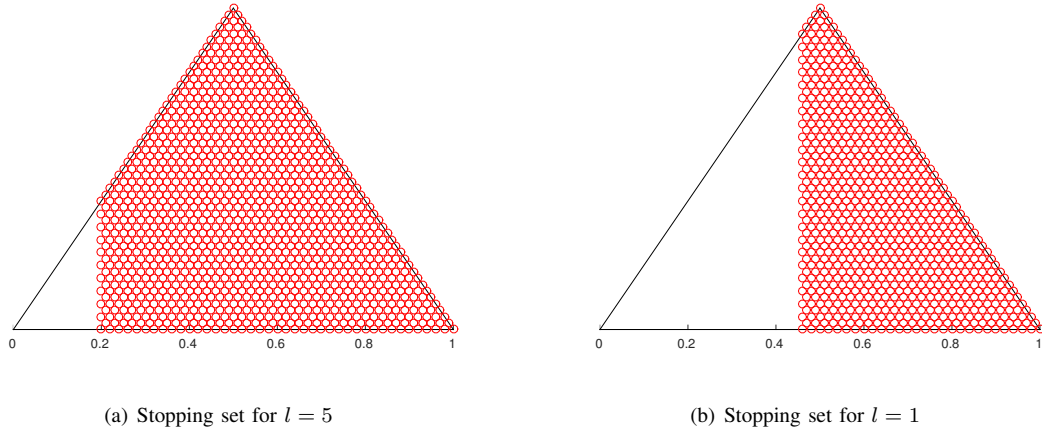


Fig. 3. Stopping set (shown in red) obtained by solving the dynamic programming in (8). The figure illustrates the sub-setting structure of the stopping sets, i.e. $S^{l-1} \subset S^l$, in Theorem 1.

B. Change Detection (Real Dataset)

In this section, we illustrate Theorem 1 for change detection problems. Change detection problems are special case of multiple stopping problems, when $L = 1$. We apply Theorem 1 for detecting changes in ground truth using an online search dataset. Online search is linked to advertising in television and online social media [34]. In addition, detecting changes in ground truth using online search data have been used for detection of outbreak of illness, political election, or major sporting events [35]. Hence, detection of changes in ground truth is important for optimizing advertising strategy. The dataset that we use is the Tech Buzz dataset from Yahoo!. We first describe the Tech Buzz dataset in Sec. V-B1 and then show through simulations that the policy obtained by solving the dynamic programming equation in (8) can be used for detecting changes in ground truth using data from online search.

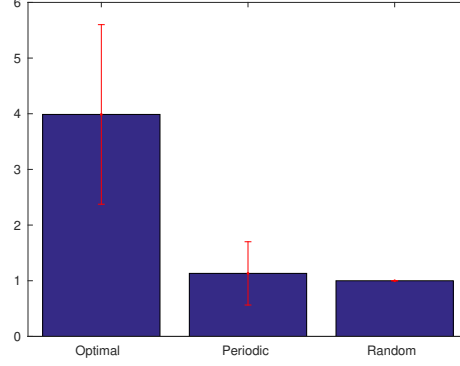


Fig. 4. Comparison between the various scheduling policies: The performance of the policy obtained by solving the dynamic programming equation in (8) is shown by ‘Optimal’. The optimal policy outperforms the conventional periodic scheduling (shown as ‘Periodic’). The random scheduling (shown as ‘Random’) is used for benchmarking the various scheduling policies.

1) *Dataset*: The dataset that we use in our study is the Yahoo! Buzz Game Transactions from the Webscope datasets¹⁶ available from Yahoo! Labs. In 2005, Yahoo! along with O’Reilly Media started a fantasy market where the trending technologies at that point were pitted against each other. For example, in the browser market there were “Internet Explorer”, “Firefox”, “Opera”, “Mozilla”, “Camino”, “Konqueror”, and “Safari”. The players in the game have access to the “buzz”, which is the online search index, measured by the number of people searching on the Yahoo! search engine for the technology. The objective of the game is to use the buzz and trade stocks accordingly.

2) *Change Detection*: We consider a subset of the data containing the WIMAX buzz scores and the number of stocks traded (volume of the stocks). The unknown valuation of the WIMAX technology is modelled using a 2–state Markov chain (“1” for high valuation and “2” for low valuation). The valuation of the stock is not observed directly, but through noisy observations on the volume of the stocks traded. Fig. 5 shows the volume of the stocks traded and the buzz during the month of April. The volume of stocks traded depend on the unknown valuation and, for ease of analysis, is quantized into 3 states (“High”, “Medium” and “Low”), denoted by 1, 2 and 3 respectively. Given the time series of the (quantized) volume of stocks traded, we obtain the hidden Markov model constituting of the transition matrix of the WIMAX valuation and the observation probability of the volume of the stocks traded given the WIMAX valuation using an EM algorithm [27]. The parameters of the Markov chain obtained using an EM algorithm are as below:

$$P = \begin{bmatrix} 1 & 0 \\ 0.1462 & 0.8538 \end{bmatrix} \quad (32)$$

$$B = \begin{bmatrix} 0.1489 & 0.4467 & 0.4044 \\ 0.3727 & 0.5325 & 0.0947 \end{bmatrix} \quad (33)$$

$$r = \begin{bmatrix} 10 & 1 \end{bmatrix} \quad (34)$$

As can be seen from Fig. 5 the WIMAX buzz and the volume of stocks traded is initially low. Hence, the objective is to find the time point at which the WIMAX value switches to the high value. The reward structure in (34) reflects the fact that choosing to *Stop* at the high value state, an agent obtains more money by trading the high value WIMAX stocks. The policy obtained by solving the dynamic programming in (8) shows a high valuation at April 18. The change point

¹⁶Yahoo! Webscope dataset: A2 - Yahoo! Buzz Game Transactions with Buzz Scores, version 1.0 http://research.yahoo.com/Academic_Relations

corresponds to Intel’s announcement of WIMAX chip¹⁷. The high valuation of WIMAX stock can also be noticed from the “spike” in the buzz around April 18 in Fig. 5.

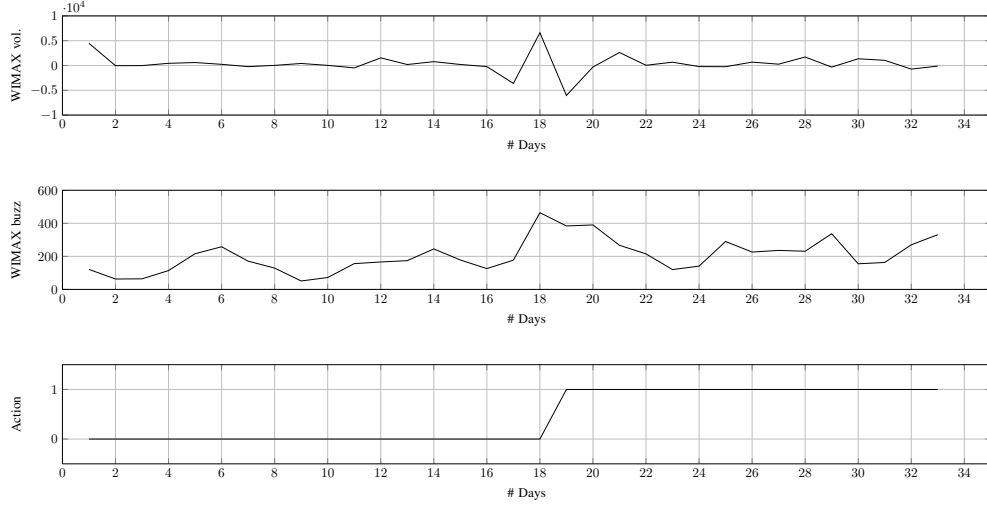


Fig. 5. The buzz scores and the trade volume for the WIMAX stock. The policy obtained through solving the dynamic programming in (8) shows a high valuation during April 18. This corresponds to Intel announcement of the WIMAX chip. The increase in valuation can be seen by a corresponding spike in the WIMAX buzz scores.

C. Ad Scheduling on Live Channels (YouTube and Twitch Datasets)

In this section, we illustrate the policy from Algorithm 1 on real data from YouTube and Twitch. In comparison to Sec. V-A and Sec. V-B, real data from YouTube and Twitch has a wide range of viewer engagement states and hence requires more states in the Markov chain model. As the number of states increases, solving the dynamic programming equation in (8) becomes impractical. Hence, we resort to best linear threshold policy through Algorithm 1. We first describe the dataset in Sec. V-C1 and then show that the policy obtained from Algorithm 1 outperforms conventional periodic scheduling.

1) *Dataset*: In this paper, we use the dataset in [36]¹⁸. The dataset consists of live session on the two popular live broadcasting platforms: YouTube Live and Twitch, between January and April 2014. The dataset contains samples of the live sessions sampled at a 5-minute interval on each of the platforms. Each sample contains the identification of the channel, the number of viewers and some additional meta-data of the channel. The main finding in [36] is that the viewer engagement is more heterogeneous than in other user-generated content platforms such as YouTube. The heterogeneity of viewer engagement in live channel can be used to opportunistically schedule advertisements.

2) *Entertainment (YouTube Live)*: In this section we use real data from YouTube Live channel and show that the policy obtained from Algorithm 1 outperforms conventional periodic scheduling.

We selected data from the “entertainment” category from the YouTube Live dataset. The channel contains data from January 01, 2014 to Jan 31, 2014, i.e. for a period of one month. Fig. 6(a) shows the distribution of the viewers of the channel during Jan 01, 2014. The parameters of the channel were obtained by the EM-Algorithm A.2.3 in [37]. The EM-

¹⁷<http://www.dailywireless.org/2005/04/17/intel-shipping-wimax-silicon/>

¹⁸The dataset is available from <http://dash.ipv6.enstb.fr/dataset/live-sessions/>

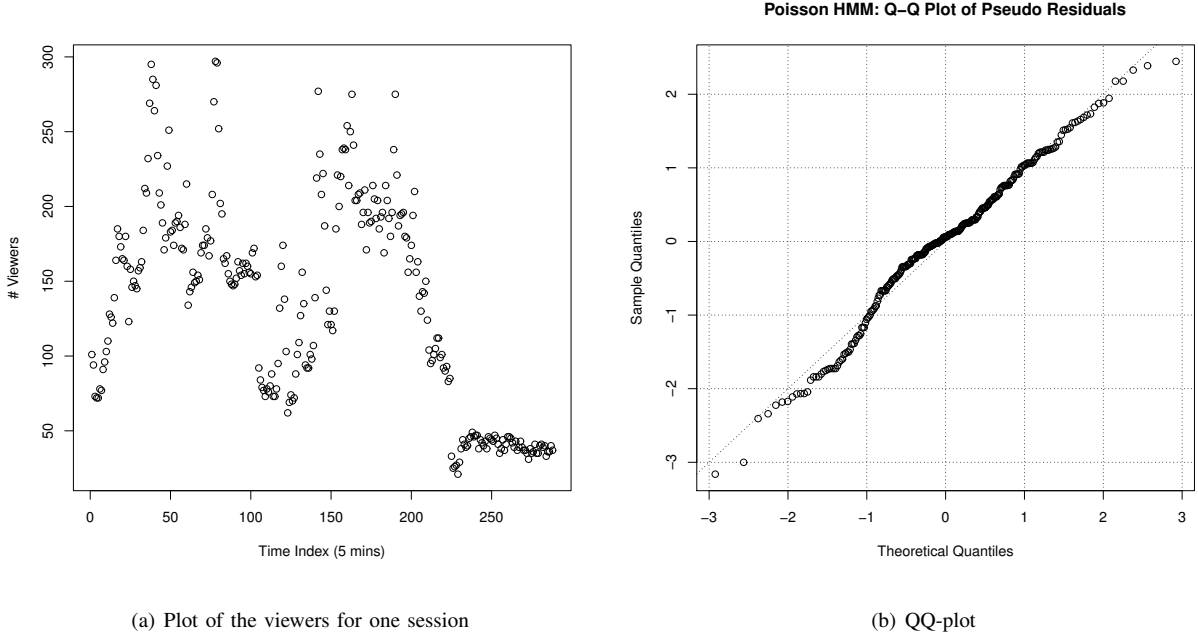


Fig. 6. Fig. 6(a) shows a plot of viewers of YouTube Live channel for one session. The distribution of the viewers is modelled by a hidden Markov process with state dependent Poisson observation process. The parameters of the model are given by (35) and (36). The QQ-plot for validating the goodness of fit is given in Fig. 6(b).

Algorithm was run for Markov Chain with 2 – 12 states. Using the AIC and BIC criteria, we selected that the channel be modelled by a 5 state Markov chain with the transition matrix in (35) and observations following state dependent Poisson distribution with mean given by (36). As can be seen from (35) the transition matrix is a first-order Markov chain validating our initial assumptions. Moreover, the transition matrix is diagonally dominant entries ensuring the TP2 assumption. This diagonally dominant entries in the transition matrix in (35) models the fact the viewer engagement of the channel changes at a slower time scale compared to the decision epochs (or sampling epochs). The reward depend on both the g in (36) and the completion and click rate α_i . Since the click rate α_i is not available in the dataset, we assume $\alpha_i = \alpha$. Due to the ordinality of the reward, α is assumed to be equal to 1. The model is validated using the QQ-plot of pseudo-residuals defined in Sec. 6.1 in [37] and is show in Fig. 6(b). As can be seen from Fig. 6(b), the QQ-plot closely follows the straight line which indicates that the model is a good fit for the data.

$$P = \begin{bmatrix} 0.94 & 0.06 & 0.00 & 0.00 & 0.00 \\ 0.02 & 0.94 & 0.04 & 0.00 & 0.00 \\ 0.00 & 0.02 & 0.96 & 0.02 & 0.00 \\ 0.00 & 0.00 & 0.06 & 0.91 & 0.03 \\ 0.00 & 0.00 & 0.00 & 0.01 & 0.99 \end{bmatrix} \quad (35)$$

$$g = [184 \quad 139 \quad 102 \quad 66 \quad 37] \quad (36)$$

Fig. 8(a) shows the comparison between the various schemes. It can be seen that the policy obtained through Algorithm 1 outperforms conventional periodic scheduling by 30%.

3) *Gaming (Twitch)*: The Twitch dataset contains channels with “gaming” content. Fig. 7(a) shows the distribution of the viewers of the channel during Jan 01, 2014. Similar, to the Sec. V-C2 above, we use the EM-Algorithm in [37] to estimate the parameters of the Markov model. The parameters of the Markov model consisting of the transition matrix

and the state dependent mean of the Poisson distribution is as given in (37) and (38), respectively. The model is validated using QQ-plot of pseudo-residuals and is shown in Fig. 7(b).

$$P = \begin{bmatrix} 0.97 & 0.03 & 0.00 & 0.00 & 0.00 \\ 0.01 & 0.96 & 0.03 & 0.00 & 0.00 \\ 0.00 & 0.02 & 0.95 & 0.03 & 0.00 \\ 0.00 & 0.00 & 0.02 & 0.96 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.02 & 0.98 \end{bmatrix} \quad (37)$$

$$g = \begin{bmatrix} 55.24 & 42.40 & 34.65 & 28.30 & 20.6 \end{bmatrix} \quad (38)$$

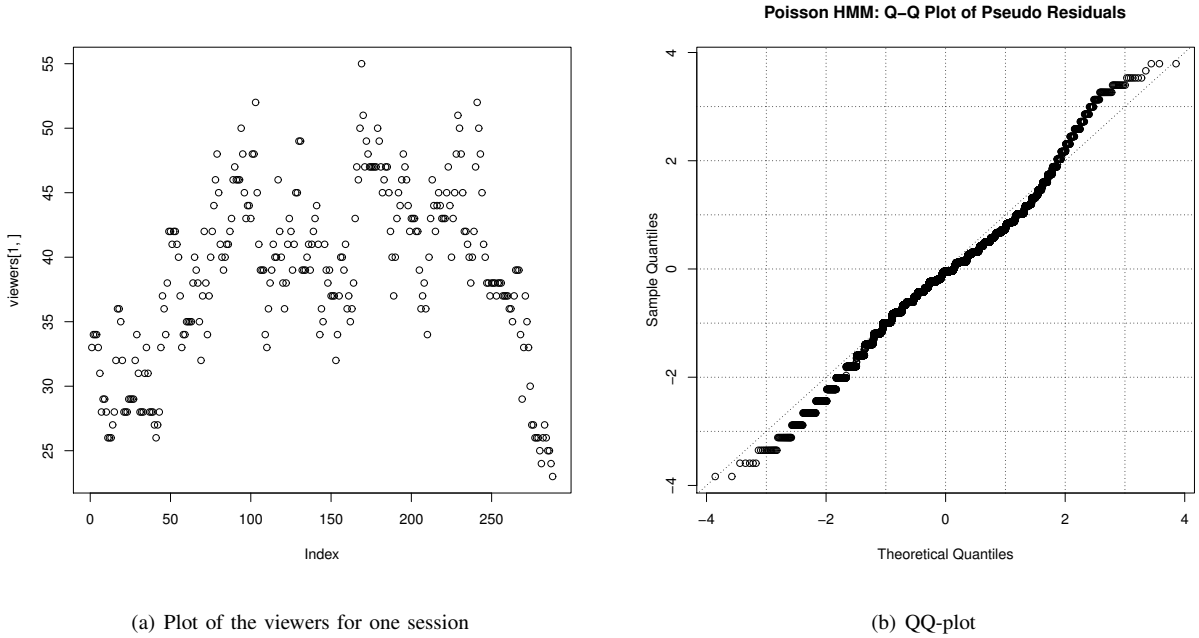


Fig. 7. Fig. 7(a) shows a plot of viewers of Twitch channel for one session. The distribution of the viewers is modelled by a hidden Markov process with state dependent Poisson observation process. The parameters of the model are given by (37) and (38). The QQ-plot for validating the goodness of fit is given in Fig. 7(b).

Fig. 8(b) shows the comparison between the various schemes. Similar to the result in the YouTube Live session, the policy obtained through Algorithm 1 outperforms conventional periodic scheduling by close to 20%. Comparing Fig. 8(a) and Fig. 8(b) we find that the performance of Algorithm 1 is lower in the Twitch channel compared to YouTube Live. This is due to the fact that Twitch is a subscription based service and the viewers are more “loyal” and hence their is less variation in the number of viewers.

VI. CONCLUSION

In this paper, we considered the problem of optimal scheduling of ads on live online social broadcasting channels. First, we cast the problem as an optimal multiple stopping problem in the POMDP framework. Second, we characterized the structural results of the optimal ad scheduling policy. Using the structural results of the optimal ad scheduling policy we computed best approximate policies using stochastic approximation. Finally, we validated the results on real datasets. First, we illustrated the analysis using synthetic data. In addition, using synthetic data, we showed that the optimal ad scheduling policy outperforms conventional scheduling techniques. Second, we show through simulations, that the policy

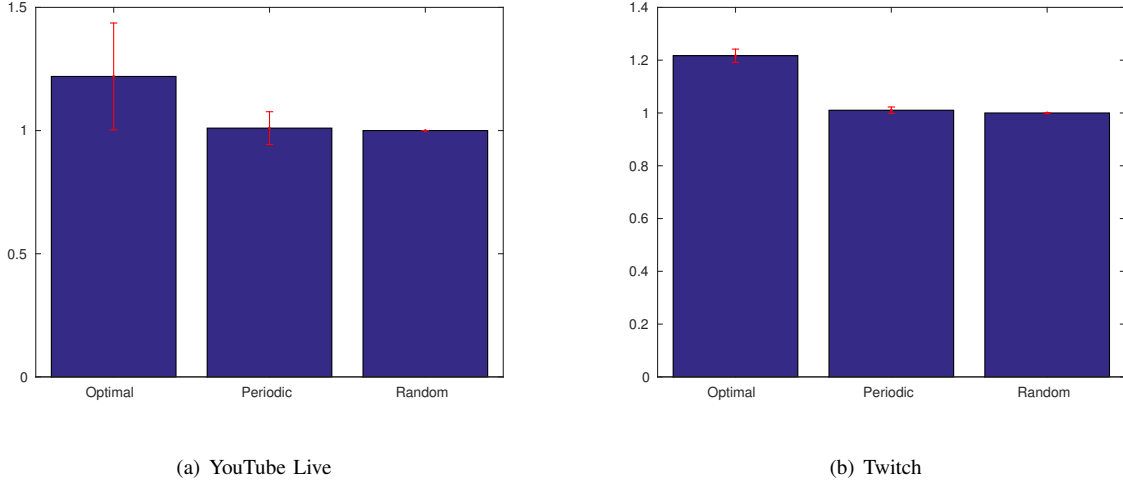


Fig. 8. Comparison of the various scheduling policies for the YouTube Live (Fig. 8(a)) and Twitch (Fig. 8(b)). The performance of the policy obtained by solving the stochastic approximation algorithm in Algorithm 1 is shown as ‘Optimal’. The policy obtained through Algorithm 1 outperforms conventional periodic scheduling (shown as ‘Periodic’) by 20%. The random scheduling (shown as ‘Random’) is used for benchmarking.

obtained from the multiple stopping problem framework, used for ad scheduling, can be used to detect changes in the ground truth using data from online search. Detecting changes in ground truth is useful for optimizing ad strategy. Finally, we show through simulations, that the best approximate ad scheduling policies obtained through stochastic approximation outperforms conventional periodic scheduling by 20 – 30%.

Extension of the current work could involve developing upper and lower myopic bounds to the optimal policy as in [38], optimizing the ad length and constraints on ad placement. These issues promise to offer interesting avenues for future work.

APPENDIX

A. First-order stochastic dominance

Definition 1: Let $\pi_1 \in \Pi$ and $\pi_2 \in \Pi$ be two belief state vectors. Then, π_1 is greater than π_2 with respect to first-order stochastic dominance—denoted as $\pi_1 \geq_s \pi_2$, if

$$\sum_{i=j}^S \pi_1(i) \leq \sum_{i=j}^S \pi_2(i) \quad \forall j \in \{1, 2, \dots, S\} \quad (39)$$

B. MLR ordering

Definition 2: Let $\pi_1 \in \Pi$ and $\pi_2 \in \Pi$ be two belief state vectors. Then, π_1 is greater than π_2 with respect to Monotone Likelihood Ratio (MLR) ordering—denoted as $\pi_1 \geq_r \pi_2$, if

$$\pi_1(j)\pi_2(i) \leq \pi_2(j)\pi_1(i), \quad i < j, \quad i, j \in \{1, \dots, S\} \quad (40)$$

MLR ordering over Π is a strong condition. In order to show threshold structure, we define the following weaker notion of MLR ordering over two types of lines.

C. MLR ordering over lines

First, we define \mathcal{H} as the $S - 2$ dimensional linear hyperplane which connects the vertices e_2, \dots, e_S as follows:

$$\mathcal{H} = \{\bar{\pi} : \bar{\pi} \in \mathcal{H} \text{ and } \bar{\pi}(1) = 0\}. \quad (41)$$

Figure 1 illustrates the definition (41) for an optimal multiple stopping problem with $S = 3$. Next, we construct two types of lines as follows:

- $\mathcal{L}(e_1, \bar{\pi})$: For any $\bar{\pi} \in \mathcal{H}$, construct the line $\mathcal{L}(e_1, \bar{\pi})$ that connects $\bar{\pi}$ to e_1 as below:

$$\mathcal{L}(e_1, \bar{\pi}) = \{\pi \in \Pi : \pi = (1 - \gamma)\bar{\pi} + \gamma e_1, 0 \leq \gamma \leq 1\}, \bar{\pi} \in \mathcal{H} \quad (42)$$

- $\mathcal{L}(e_S, \bar{\pi})$: For any $\bar{\pi} \in \mathcal{H}$, construct the line $\mathcal{L}(e_S, \bar{\pi})$ that connects $\bar{\pi}$ to e_S as below:

$$\mathcal{L}(e_S, \bar{\pi}) = \{\pi \in \Pi : \pi = (1 - \gamma)\bar{\pi} + \gamma e_S, 0 \leq \gamma \leq 1\}, \bar{\pi} \in \mathcal{H} \quad (43)$$

With an abuse of notation, we denote $\mathcal{L}(e_1, \bar{\pi})$ by $\mathcal{L}(e_1)$ and $\mathcal{L}(e_S, \bar{\pi})$ by $\mathcal{L}(e_S)$. Figure 1 illustrates the definition of $\mathcal{L}(e_1)$.

Definition 3 (MLR ordering on lines): π_1 is greater than π_2 with respect to MLR ordering on the lines $\mathcal{L}(e_1)$, denoted as $\pi_1 \geq_{\mathcal{L}_1} \pi_2$, if $\pi_1, \pi_2 \in \mathcal{L}(e_1, \bar{\pi})$, for some $\bar{\pi} \in \mathcal{H}$ and $\pi_1 \geq_r \pi_2$.

The MLR order is a partial order, however, the MLR ordering on lines is a complete order. The MLR on lines requires less stringent conditions and can be used for devising threshold policies over lines.

D. TP2 ordering

Definition 4 (TP2 ordering): A transition probability matrix, A is Totally Positive of order 2 (TP2), if all the second order minors are non-negative i.e. the determinants

$$\begin{vmatrix} a_{i_1 j_1} & a_{i_1 j_2} \\ a_{i_2 j_1} & a_{i_2 j_2} \end{vmatrix} \geq 0, \forall i_2 \geq i_1, j_2 \geq j_1 \quad (44)$$

An important consequence of the TP2 ordering in Definition 4 is the following theorems, which states that the filter $T(\pi, y)$ preserves MLR dominance.

Theorem 4 (Theorem 10.3.1 in [27]): If the transition matrix, P , and the observation matrix, B , satisfies the condition in III-B1 and III-B1, then

- For $\pi_1 \geq_r \pi_2$, the filter satisfies $T(\pi_1, y) \geq_r T(\pi_2, y)$.
- For $\pi_1 \geq_r \pi_2$, $\sigma(\pi_1, y) \geq_s \sigma(\pi_2, y)$

Theorem 5 ([39]): $\pi_1 \geq_s \pi_2$ if and only if for any increasing function $\phi(\cdot)$, $\mathbb{E}_{\pi_1} \{\phi(x)\} \geq \mathbb{E}_{\pi_2} \{\phi(x)\}$

To prove Prop. 1, Prop. 2 and Prop. 3, we assume that the proposition hold for all values less than k .

E. Proof of Prop. 1

Recall from (10),

$$V_k(\pi, l) = \max_{u \in \{1, 2\}} Q_k(\pi, l, u),$$

To prove Prop. 1, we show $Q_k(\pi, l, u)$ is increasing in π for $u = \{1, 2\}$.

Recall from (12),

$$Q_k(\pi, l, 1) = r'\pi + \rho \sum_y V_{k-1}(T(\pi, y), l-1)\sigma(\pi, y),$$

Using Theorem 4, Theorem 5 and the induction hypothesis, the term $\sum_y V_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)$ is increasing in π . From Assumption III-B1, $r'\pi$ is increasing in π . The proof for $Q_k(\pi, l, 2)$ increasing in π is similar and is omitted. Hence, $V_k(\pi, l)$ is increasing in π . ■

F. Proof of Prop. 2

The proof follows by induction. Recall from (19), we have

$$W_k(\pi, l-1) = \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)\mathcal{I}_{C_k^{l-1}}(\pi) + r'\pi\mathcal{I}_{C_k^{l-2} \cap S_k^{l-1}}(\pi) + \sum_y W_{k-1}(T(\pi, y), l-2)\sigma(\pi, y)\mathcal{I}_{S_k^{l-2}}(\pi) \quad (45)$$

Hence, we compare $W_k(\pi, l)$ and $W_k(\pi, l-1)$ in the following 4 regions:

a.) S_k^{l-2} :

$$W_k(\pi, l) - W_k(\pi, l-1) = \sum_y (W_{k-1}(T(\pi, y), l-1) - W_{k-1}(T(\pi, y), l-2))\sigma(\pi, y),$$

which is non-negative by the induction assumption.

b.) $C_k^{l-2} \cap S_k^{l-1}$:

$$W_k(\pi, l) - W_k(\pi, l-1) = \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y) - r'\pi,$$

which is non-negative since $\pi \in S_k^{l-1}$.

c.) $C_k^{l-1} \cap S_k^l$:

$$W_k(\pi, l) - W_k(\pi, l-1) = r'\pi - \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y),$$

which is non-negative since $\pi \in C_k^{l-1}$.

d.) C_k^l :

$$W_k(\pi, l) - W_k(\pi, l-1) = \sum_y (W_{k-1}(T(\pi, y), l) - W_{k-1}(T(\pi, y), l-1))\sigma(\pi, y),$$

which is non-negative by the induction assumption.

■

G. Proof of Prop. 3

If $\pi \in S_k^{l-1}$, then $r'\pi \geq \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)$. By Prop. 2, $r'\pi \geq \sum_y W_{k-1}(T(\pi, y), l)\sigma(\pi, y)$. Hence $\pi \in S_k^l$. ■

H. Proof of Theorem 1

Existence of optimal policy: In order to show the existence of a threshold policy of \mathcal{L}_1 , we need to show that $Q_{k+1}(\pi, l, 2) - Q_{k+1}(\pi, l, 1)$ is supermodular in $\pi \in \mathcal{L}(e_1, \bar{\pi})$. Since,

$$Q_{k+1}(\pi, l, 2) - Q_{k+1}(\pi, l, 1) = \rho \sum_y W_k(T(\pi, y), l)\sigma(\pi, y) - r'\pi.$$

We need to show that $\rho \sum_y W_k(T(\pi, y), l) \sigma(\pi, y) - r' \pi$ is decreasing in π .

$$\begin{aligned}
\rho \sum_y W_k(T(\pi, y), l) \sigma(\pi, y) - r' \pi &= \sum_y (\rho W_k(T(\pi, y), l) - r' \pi) \sigma(\pi, y) \\
&= \sum_y ((\rho W_k(T(\pi, y), l) - \rho r' T(\pi, y)) - (r' \pi - \rho r' T(\pi, y))) \sigma(\pi, y) \\
&= \rho \sum_y (W_k(T(\pi, y), l) - r' T(\pi, y)) \sigma(\pi, y) - r' (I - \rho P') \pi
\end{aligned} \tag{46}$$

The term $r' (I - \rho P') \pi$ in (46) is decreasing in π due to our assumption. Hence, to show that $\rho \sum_y W_k(T(\pi, y), l) \sigma(\pi, y) - r' \pi$ is decreasing in π it is sufficient to show that $W_k(\pi, l) - r' \pi$ is decreasing in π . Define,

$$\bar{W}_k(\pi, l) \triangleq W_k(\pi, l) - r' \pi \tag{47}$$

Now, $\bar{W}_k(\pi, l) =$

$$\begin{aligned}
&\left(\sum_y \rho ((\bar{W}_{k-1}(T(\pi, y), l) + r' T(\pi, y)) - r' \pi) \sigma(\pi, y) \right) \mathcal{I}_{C_k^l}(\pi) + \left(\sum_y \rho ((\bar{W}_{k-1}(T(\pi, y), l-1) + r' T(\pi, y)) - r' \pi) \sigma(\pi, y) \right) \mathcal{I}_{S_k^l}(\pi) \\
&= \left(\sum_y (\rho \bar{W}_{k-1}(T(\pi, y), l) \sigma(\pi, y)) - r' (I - \rho P)' \pi \right) \mathcal{I}_{C_k^l}(\pi) + \left(\sum_y (\rho \bar{W}_{k-1}(T(\pi, y), l-1) \sigma(\pi, y)) - r' (I - \rho P)' \pi \right) \mathcal{I}_{S_k^l}(\pi)
\end{aligned} \tag{48}$$

We prove using induction that $\bar{W}_k(\pi, l)$ is decreasing in π , using the recursive relation over k in (48).

For $k = 0$,

$$\bar{W}_0(\pi, l) = W_0(\pi, l) - r' \pi = V_0(\pi, l) - V_0(\pi, l-1) - r' \pi \tag{49}$$

The initial conditions of the value iteration algorithm can be chosen such that $\bar{W}_0(\pi, l)$ in (49) is decreasing in π . A suitable choice of the initial conditions is given below:

$$V_0(\pi, l) = r' \left(\sum_{j=0}^{l-1} \rho^j P^j \right)' \pi. \tag{50}$$

The intuition behind the initial conditions in (50) is that the value function, $V_0(\pi, l)$ gives the expected total reward if we stop l times successively starting at belief π . Hence, it is clear that $\bar{W}_k(\pi, l)$ is decreasing in π , if $\bar{W}_{k-1}(\pi, l)$ is decreasing in π , finishing the induction step.

Characterization of the switching curve Γ_l : For each $\bar{\pi} \in \mathcal{H}$ construct the line segment $\mathcal{L}(e_1, \bar{\pi})$. The line segment can be described as $(1 - \varepsilon) \bar{\pi} + \varepsilon e_1$. On the line segment $\mathcal{L}(e_1, \bar{\pi})$ all the belief states are MLR orderable. Since $\mu^*(\pi, l)$ is monotone decreasing in π , for each l , we pick the largest ε such that $\mu^*(\pi, l) = 1$. The belief state, $\pi^{\varepsilon^*, \bar{\pi}}$ is the threshold belief state, where $\varepsilon^* = \inf \{ \varepsilon \in [0, 1] : \mu^*(\pi^{\varepsilon, \bar{\pi}}, l) = 1 \}$. Denote by $\Gamma(\bar{\pi}) = \pi^{\varepsilon^*, \bar{\pi}}$. The above construction implies that there is a unique threshold $\Gamma(\bar{\pi})$ on $\mathcal{L}(e_1, \bar{\pi})$. The entire simplex can be covered by considering all pairs of lines $\mathcal{L}(e_1, \bar{\pi})$, for $\bar{\pi} \in \mathcal{H}$, i.e. $\Pi(X) = \cup_{\bar{\pi} \in \mathcal{H}} \mathcal{L}(e_1, \bar{\pi})$. Combining, all points yield a unique threshold curve in $\Pi(X)$ given by $\Gamma = \cup_{\bar{\pi} \in \mathcal{H}} \Gamma(\bar{\pi})$.

Connectedness of S^l : Since $e_1 \in S^l$ for all l , call S_a^l , the subset of S^l that contains e_1 . Suppose S_b^l is the subset that was disconnected from S_a^l . Since every point on $\Pi(X)$ lies on the line segment $\mathcal{L}(e_1, \bar{\pi})$, for some $\bar{\pi}$, there exists a line segment starting from $e_1 \in S_a^l$ that would leave the region S_a^l , pass through the region where action 2 is optimal and then intersect region S_b^l , where action 1 is optimal. But, this violates the requirement that the policy $\mu^*(\pi, l)$ is monotone on $\mathcal{L}(e_1, \bar{\pi})$. Hence, S_a^l and S_b^l are connected.

Connectedness of C^l : Assume $e_X \in C^l$, otherwise $C^l = \phi$ and there is nothing to prove. Call the region that contains e_X as C_a^l . Suppose $C_b^l \subset C^l$ is disconnected from C_a^l . Since every point in $\Pi(X)$ can be covered by a line segment $\mathcal{L}(e_X, \bar{\pi})$, for some $\bar{\pi}$. Then, there exists a line starting from $e_X \in C_a^l$ would leave region C_a^l , pass through the region where action 1 is optimal and then intersect the region C_b^l (where action 2 is optimal). But this violates the monotone property of $\mu^*(\pi, l)$.

Sub-setting structure: The proof is straightforward from Prop. 3. ■

I. Proof of Theorem 3

For $l_1 > l_2$, due to the sub-setting structure in Theorem 1 $S^{l_2} \subset S^{l_1}$. This implies the following

$$\begin{aligned} \mu_\theta(l_2, \pi) &\geq \mu_\theta(l_1, \pi) \\ \begin{bmatrix} 0 & 1 & \theta_{l_2} \end{bmatrix} \begin{bmatrix} \pi \\ -1 \end{bmatrix} &\geq \begin{bmatrix} 0 & 1 & \theta_{l_1} \end{bmatrix} \begin{bmatrix} \pi \\ -1 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & \theta_{l_2} - \theta_{l_1} \end{bmatrix} \begin{bmatrix} \pi \\ -1 \end{bmatrix} &\geq 0 \end{aligned} \quad (51)$$

It is straightforward to check that the conditions in (22) in Theorem 3 satisfy the conditions in (51). ■

REFERENCES

- [1] T. Smith, M. Obrist, and P. Wright, "Live-streaming changes the (video) game," in *Proc. of the 11th European Conference on Interactive TV and Video*. ACM, 2013, pp. 131–138.
- [2] S. Bollapragada, M. R. Bussieck, and S. Mallik, "Scheduling commercial videotapes in broadcast television," *Oper. Res.*, vol. 52, no. 5, pp. 679–689, Oct. 2004.
- [3] D. G. Popescu and P. Crama, "Ad revenue optimization in live broadcasting," *Management Science*, vol. 62, no. 4, pp. 1145–1164, 2015.
- [4] S. Seshadri, S. Subramanian, and S. Souyris, "Scheduling spots on television," 2015.
- [5] H. Kang and M. P. McAllister, "Selling you and your clicks: examining the audience commodification of google," *Journal for a Global Sustainable Information Society*, vol. 9, no. 2, pp. 141–153, 2011.
- [6] R. Terlutter and M. L. Capella, "The gamification of advertising: analysis and research directions of in-game advertising, advergames, and advertising in social network games," *Journal of Advertising*, vol. 42, no. 2-3, pp. 95–112, 2013.
- [7] J. Turner, A. Scheller-Wolf, and S. Tayur, "Scheduling of dynamic in-game advertising," *Operations Research*, vol. 59, no. 1, pp. 1–16, 2011.
- [8] N. Archak, V. Mirrokni, and S. Muthukrishnan, "Budget optimization for online campaigns with positive carryover effects," in *Proc. of the 8th International Conference on Internet and Network Economics*. Springer-Verlag, 2012, pp. 86–99.
- [9] N. Archak, V. S. Mirrokni, and S. Muthukrishnan, "Mining advertiser-specific user behavior using adfactors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. ACM, 2010, pp. 31–40.
- [10] T. Nakai, "The problem of optimal stopping in a partially observable markov chain," *Journal of optimization Theory and Applications*, vol. 45, no. 3, pp. 425–442, 1985.
- [11] W. Stadje, "An optimal k-stopping problem for the poisson process," in *Mathematical Statistics and Probability Theory*. Springer, 1987, pp. 231–244.
- [12] M. Nikolaev, "On optimal multiple stopping of markov sequences," *Theory of Probability & Its Applications*, vol. 43, no. 2, pp. 298–306, 1999.
- [13] A. Krasnosielska-Kobos, "Multiple-stopping problems with random horizon," *Optimization*, vol. 64, no. 7, pp. 1625–1645, 2015.
- [14] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995, vol. 1, no. 2.
- [15] E. Bayraktar and R. Kravitz, "Quickest detection with discretely controlled observations," *Sequential Analysis*, vol. 34, no. 1, pp. 77–133, 2015.
- [16] J. Geng, E. Bayraktar, and L. Lai, "Bayesian quickest change-point detection with sampling right constraints," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6474–6490, 2014.
- [17] T. L. Lai, "On optimal stopping problems in sequential hypothesis testing," *Statistica Sinica*, vol. 7, no. 1, pp. 33–51, 1997.

- [18] —, *Sequential analysis*. Wiley Online Library, 2001.
- [19] S. H. J. Alexander G. Nikolaev, “Stochastic sequential decision-making with a random number of jobs,” *Operations Research*, vol. 58, no. 4, pp. 1023–1027, 2010.
- [20] S. Savin and C. Terwiesch, “Optimal product launch times in a duopoly: Balancing life-cycle revenues with product cost,” *Operations Research*, vol. 53, no. 1, pp. 26–47, 2005.
- [21] I. Lobel, J. Patel, G. Vulcano, and J. Zhang, “Optimizing product launches in the presence of strategic consumers,” *Management Science*, vol. 62, no. 6, pp. 1778–1799, 2015.
- [22] K. E. Wilson, R. Szechtman, and M. P. Atkinson, “A sequential perspective on searching for static targets,” *European Journal of Operational Research*, vol. 215, no. 1, pp. 218 – 226, 2011.
- [23] M. Atkinson, M. Kress, and R.-J. Lange, “When is information sufficient for action? search with unreliable yet informative intelligence,” *Operations Research*, vol. 64, no. 2, pp. 315–328, 2016.
- [24] I. D. Askwith, “Television 2.0: Reconceptualizing tv as an engagement medium,” Ph.D. dissertation, Massachusetts Institute of Technology, 2007.
- [25] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, “Understanding user behavior in large-scale video-on-demand systems,” *SIGOPS Oper. Syst. Rev.*, vol. 40, no. 4, pp. 333–344, Apr. 2006.
- [26] V. Krishnamurthy and D. V. Djonin, “Structured threshold policies for dynamic sensor scheduling—a partially observed markov decision process approach,” *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 4938–4957, Oct 2007.
- [27] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.
- [28] —, “How to schedule measurements of a noisy Markov chain in decision making?” *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4440–4461, July 2013.
- [29] —, “Bayesian sequential detection with phase-distributed change time and nonlinear penalty – A POMDP Lattice programming approach,” *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 7096–7124, Oct 2011.
- [30] J. J. M. Eric Johnson, “Infinitesimal perturbation analysis: A tool for simulation,” *The Journal of the Operational Research Society*, vol. 40, no. 3, pp. 243–254, 1989.
- [31] G. C. Pflug, *Optimization of stochastic models: the interface between simulation and optimization*. Springer Science & Business Media, 2012, vol. 373.
- [32] J. C. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005, vol. 65.
- [33] I.-J. Wang and J. C. Spall, “Stochastic optimisation with inequality constraints using simultaneous perturbations and penalty functions,” *International Journal of Control*, vol. 81, no. 8, pp. 1232–1238, 2008.
- [34] M. Joo, K. C. Wilbur, B. Cowgill, and Y. Zhu, “Television advertising and online search,” *Management Science*, vol. 60, no. 1, pp. 56–73, 2013.
- [35] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.
- [36] K. Pires and G. Simon, “Youtube live and twitch: A tour of user-generated live streaming systems,” in *Proceedings of the 6th ACM Multimedia Systems Conference*, ser. MMSys ’15. ACM, 2015, pp. 225–230.
- [37] W. Zucchini and I. L. MacDonald, *Hidden Markov models for time series: an introduction using R*. CRC press, 2009.
- [38] V. Krishnamurthy and U. Pareek, “Myopic bounds for optimal policy of POMDPs: An extension of Lovejoy’s structural results,” *Operations Research*, vol. 62, no. 2, pp. 428–434, 2015.
- [39] P. C. Kiessler, “Comparison methods for stochastic models and risks,” *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 704–704, 2005.